Sample Size Calculation For Complex Sampling Designs (Version 1.0)

Felderer, Barbara; Sand, Matthias; Bruch, Christian

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Felderer, B., Sand, M., & Bruch, C. (2022). *Sample Size Calculation For Complex Sampling Designs (Version 1.0).* (GESIS Survey Guidelines). Mannheim: GESIS - Leibniz-Institut für Sozialwissenschaften. https://doi.org/10.15465/gesis-sg_en_042

gesis Leibniz Institute for the Social Sciences

Sample Size Calculation For Complex Sampling Designs

Barbara Felderer, Matthias Sand, & Christian Bruch

February 2022, Version 1.0

Abstract

Before conducting a survey, researchers frequently ask themselves how large the resulting sample of respondents needs to be to answer their research questions. In this guideline, we discuss how sample size calculation is affected by the sampling design. We give practical advice on how to conduct sample size calculation for complex samples.

Citation

Felderer, Barbara, Sand, Matthias, & Bruch, Christian (2022). Sample Size Calculation For Complex Sampling Designs. Mannheim, GESIS - Leibniz Institute for the Social Sciences (GESIS- Survey Guidelines).

DOI: 10.15465/gesis-sg_en_042

1. Introduction

Before conducting a survey, researchers need to decide what sample size they need to answer their research question. The sample size needs to be large enough to be able to precisely estimate a statistic of interest for the target population and/or perform statistical tests. While more respondents are preferable for any statistical analysis, the sample size is limited by practical constraints like survey costs or time constraints (Groves, 2005).

To decide on the minimum sample size that is needed for an analysis, one must be sure which variable(s) to include in the analysis and which tests to perform. The minimum sample size might be very different depending on the survey variables and their distribution. For surveys including different survey questions, sample size analysis should be conducted for each variable and the largest minimum sample size should be chosen for the survey. The minimum sample size in general depends on the kind of analysis that is conducted (e.g., estimate a mean or proportion, conduct specific statistical tests), the desired confidence level and statistical power, the distribution of the outcome in the population. Whereas the former can be set by the researcher, the latter is fixed but in most applications unknown.

Among other, the minimum sample size depends on the sampling design. Roughly speaking, sampling designs that result in a higher variance of the point estimator will likely need a higher sample size. Sample size calculation for simple random samples (srs) has been extensively covered in the literature (see for example Gelman & Hill, 2006; Valliant, Dever, & Kreuter, 2013) and many R packages are available to perform the calculation, for example *pwr* (Champely, 2020). The very comprehensive online tool G*Power (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007) should also be mentioned. The list of tools to conduct sample size calculation for srs is by far not complete, there are too many to name them all here.

For sample size calculation of complex sampling designs[1], we refer for example to the online sample size calculator from UK sample (https://uksamples.co.uk/sample-size-calculator) or *PracTools* (Valliant, Dever, & Kreuter, 2021). These tools make use of the intra-class correlation coefficient (ICC) as a measure of similarity within an (observed) cluster and/or the design effect of a particular sample design to calculate the minimum sample size. While sample size calculation is an easy task in theory, in practice, it is often hard to decide which ICC or design effect to assume for a specific study.

In this guideline, we start by discussing the role of the ICC and the design effect when conducting sample size calculation for complex sampling designs and its effect on the minimum needed sample size as compared to srs. To guide the decision on which design effect to assume for a specific survey, we compile design effects from the ESS that readers can use as a reference. Since a specific survey is, however, not always comparable to a general purpose survey such as the ESS, we furthermore give a summary of rules of thumb that can be found in the literature to give the reader an insight on what the "bare minimum" of the sample size for a particular research question is supposed to look like.

2. Sample Size for Estimating Means and Proportions (in Complex Sampling Designs)

Let us assume we are interested in learning about a population variable Y with observation y_i for element i. The population variable is characterized by the mean ($\bar{Y}$) and variance ($var(Y)$) which are unknown. We thus want to conduct a survey to estimate the mean of y based on a sample with observations y_i

[1] By complex sampling designs we mean all types of sampling designs that are not simple random samples.

($i = 1 \ldots n$) where n is the sample size. The estimated mean is given by $\bar{y}$[2]. To account for the variability due to sampling, we want to put a confidence interval (CI) around the point estimate $\bar{y}$. In our case, we specify a confidence level of $\alpha = 0.05$.

The CI for $\bar{y}$ is given by[3]:

$$\left[\bar{y} - z_{1-\frac{\alpha}{2}}\sqrt{\hat{var}(\bar{y})}, \bar{y} + z_{1-\frac{\alpha}{2}}\sqrt{\hat{var}(\bar{y})}\right] \tag{1}$$

where $\hat{var}(\bar{y})$ is the variance of the estimator for $\bar{y}$ and $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ - quantile of the standard normal distribution.

When **simple random sampling** is applied and the finite population correction can be neglected, $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ and $\hat{var}(\bar{y}) = \frac{\hat{var}(y)}{n}$ where $\hat{var}(y) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$ is the survey based estimator for $var(y)$. For srs, the CI is given by

$$\left[\bar{y} - z_{1-\frac{\alpha}{2}}\frac{\sqrt{\hat{var}(y)}}{\sqrt{n}}, \bar{y} + z_{1-\frac{\alpha}{2}}\frac{\sqrt{\hat{var}(y)}}{\sqrt{n}}\right] \tag{2}$$

For all factors kept constant, larger n lead to higher precision of $\bar{y}$ and a smaller CI. Half the length of the CI is given by

$$e = z_{1-\frac{\alpha}{2}}\frac{\sqrt{\hat{var}(y)}}{\sqrt{n}} \tag{3}$$

The choice of the maximal width of the CI (2*e) heavily depends on the research question. Let us for example consider a researcher wants to forecast whether certain political parties will pass the 5% threshold to enter the German Bundestag. For larger parties that are expected to receive for example 20% of the votes, a broader CI can be tolerated to evaluate whether they are above 5% or not. For smaller parties, a higher precision and thus smaller CI will be needed.

The idea of sample size calculation for estimating means and proportions is to determine the minimum sample size that is needed to receive a desired precision and thus maximal width of the CI. For srs, this is simply done by solving for n (Gabler & Häder, 2015):

$$n_{srs} \geq (\frac{z_{1-\frac{\alpha}{2}}}{e})^2 \hat{var}(y). \tag{4}$$

In practice, the variance of y is not known before the survey is conducted. The choice of the variance can be informed by variances (or standard deviations) found in comparable studies or by a pretest conducted before the main study. Analyzing proportions, assuming the proportion to be 0.5 (and thus the variance to be 0.25) can be used as a worst-case scenario. For continuous variables no such worst-case bounds exist for the mean or the variance.

Sample size calculation heavily depends on the estimated variance of y which is a component of the estimated variance of $\bar{y}$. Keeping all other factors constant, a higher variance implies a higher minimum sample size to receive a certain width of the CI.

[2] The correct estmimation of the survey mean depends on the sampling design.

[3] Please note that the confidence interval for proportions is build analogously.

The above formulas (2) to (4) do only hold for srs. For more **complex sampling designs**, one way to compute the sample size is to integrate the design effect (deff) in the calculation (Lynn, Häder, & Gabler, 2006) as the variance of an estimator is affected by the complex sampling design (Gabler & Häder, 2015). In general, cluster sampling and sampling with unequal inclusion probabilities increases the variance of $\bar{y}$. Stratification, on the other side can reduce the variance of $\bar{y}$.

To determine the minimum sample size in a complex sampling design, the design elements need to be taken into account. For a complex sample, the design effect is given by:

$$deff = \frac{var_{complex}(\bar{y})}{var_{srs}(\bar{y})} \tag{5}$$

where $var_{complex}(\bar{y})$ is the variance of $\bar{y}$ under a complex sampling design and $var_{srs}(\bar{y})$ is the variance of $\bar{y}$ under simple random sample.

For clustered samples and samples with unequal inclusion probabilities the $deff$ is generally larger or equal to one ($deff \geq 1$). Stratification can lead to $deff \leq 1$. The effect of stratification is assumed to be small and is usually not accounted for when determining $deff$ (Lynn et al., 2006).

The design effect needs to be estimated and according to Gabler, Häder, & Lahiri (1999) and Kish (1987) a model based approach can be applied to account for clustering and/or unequal inclusion probabilities:

$$\begin{aligned} \widehat{deff} &= n\frac{\sum_{g=1}^{G} n_g w_g^2}{(\sum_{g=1}^{G} n_g w_g)^2} * \left[1 + (\bar{b} - 1)ICC\right] \\ &= \widehat{deff_w} * \widehat{deff_c} \end{aligned} \tag{6}$$

The design effect is the product of the design effect due to unequal inclusion probabilities ($\widehat{deff_w}$) and the design effect due to clustering ($\widehat{deff_c}$) (see Gabler et al. (1999), Kish (1987)). The n_g are the number of observations in the g^{th} weighting class with $n = \sum_{g=1}^{G} n_g$, G is the number of weighting classes, w_g are the weights in the g^{th} weighting class and $\bar{b}$ is the mean cluster size. The intra-class correlation coefficient (ICC) is a measure describing how similar the observations within the same cluster are (Gabler et al., 1999). The ICC equals 1 if all cluster members are equal but is usually found to be small for general population surveys (Kalton, Brick, & Lê, 2005). The equation illustrates the relationship between the ICC and the variance of an estimate for a particular variable. More similarity of elements within the clusters with respect to the variable of interest leads to a higher ICC and ceteris paribus to a higher variance (as compared to a simple random sample). Hence, to obtain a particular variance (and width of the confidence interval) the sample size would need to be higher as under srs with a higher ICC (Campbell, Grimshaw, & Steen, 2000).

To account for the complex survey design, the sample size is given by:

$n_{compl} = n_{srs} * \widehat{deff}$

where n_{srs} is the sample size calculated for simple random sample. Note that $\widehat{deff}$ is also usually not known before the survey has been conducted but may be estimated by a similar survey that has been conducted before.

3. Practical Advice

To determine the minimum sample size needed to achieve a specific width of the CI or to perform a statistical test we either need to know about the population parameters of interest or to make strong assumptions about them. For a complex survey design, additional assumptions need to be made about the parameters that are used to estimate deff. These parameters are in most applications hard to determine.

For the means and variances it might be helpful to consult official statistics or other surveys asking a similar question to get an idea on which values to expect. For large-scale surveys usually a field test is conducted. Information from the field-test data can be used to determine the sample size for the main study. Concerning the design effect, it is helpful to determine what design effects have been found in studies with similar sampling scheme.

3. 1 Design Effects Found in Other Surveys

Usually, the design effect is not known in practice and it can be very hard to determine which size is realistic. Information about design effects are often hard to find because researchers rarely report the design effect of their particular survey. One way to obtain such an estimation of the expected design effect could be to contact persons that already conducted a survey with a similar design (in the same country). As an alternative, a rough estimate could for example be obtained from sources such as the *European Social Survey (ESS)*. In their seventh round, the ESS reported in detail about their model based design effects approximation for several participating countries and displayed their average estimated design effects (Koen Beullens, 2014). An abbreviated version can be found in Table 1.

Countries	Frame	Stages	ICC	$deff_c$	$deff_p$	$deff$
Belgium	Official	2	0.05	1.19	1	1.19
Finnland	Official	1	0	1	1	1
France	Random Walk	4	0.04	1.4	1.2	1.68
Germany	Official	2	0.04	1.58	1.1	1.738
Portugal	Random Walk	3	0.065	1.6	1.25	2
UK	Postal	5	0.043	1.39	1.27	1.7653

Table 1: ESS Round 1: Design effects for different countries.

The estimates presented in Table 1 are only an approximation of the mean design effect for a general purpose survey. As already stated in the previous section, the actual design effect is always measured as the proportion of the variance of the estimator of a variable of interest under a complex survey design and its counterpart under simple random sample. Hence, there are two important factors to keep in mind, when searching for an approximation to plan the sample size for ones own survey:

1. the sampling design of the survey for which the sample size should be calculated
2. the (population) variance of the variable of interest (if it is a single purpose survey). In case of a multipurpose survey, one should calculate the necessary sample size for each relevant variable and decide on the largest.

Gabler, Häder, & Lynn (2006) already demonstrated for the first round of the ESS how the design effect may differ for different sampling designs (within a particular country) and different variables of interest. Table 2 displays an abbreviated version of the original results published by Gabler et al. (2006) for the German case.

Variable of interest	Design effect: $S1$	Design effect: $S2$	Design effect: $S3$
Persons in household	1.87	1.85	1.74
Years of education	3.25	2.8	2.88
Discriminated by religion	1.22	1.05	1.08
Left-right-sacle	1.7	1.65	1.58
Satisfaction with life	2.06	1.74	1.81
Religiosity	1.94	1.75	1.75
Political activism	3.26	2.83	2.89

Table 2: ESS Round 1: Three approximations of the design effects for different sampling designs.

To illustrate how the design effect differs by sampling design, the authors contrast the design effect they found for the ESS to simulated design effects that would have been realized with alternative designs. Design effect "$S1$" in Table 2 refers to a sampling design with unequal probabilities of inclusion, design effect "$S2$" to a simulated sampling design with equal inclusion probabilities within domain/stratum and design effect "$S3$" to a simulated design where the sample size within stratum is proportionally allocated to a stratum's population size that was applied in ESS round 1 for Germany. As one can see, even within the same sample design, the actual design effect differs between the variables of interest. Moreover, the differences between the variables are larger than the differences between the designs. In their original study, Gabler et al. (2006) used the second and third approximation of the design effect to showcase the potential of underestimation for the design effect when a different sampling design (than the original) is assumed. Therefore, they recommend to use the most conservative approximation that assumes unequal probabilities of inclusion on an individual level.

Since approximations of the design effect for a given country with a given sample design and a given variable of interest is often hard to come by, a general rule of thumb may also be to orient oneself on multi-purpose surveys that reported their design effects (such as the ESS) and err on the conservative side. Hence, for a survey of the German population, one could use the reported design effect of 1.74 and round it to 2 in order to obtain the required sample size. Please keep in mind that this is only a rough approximation that should be adjusted if better information is available.

3. 2. Rules of Thumb

Multiplying the required sample size under simple random sample with the (model based approximation) of the design effect of previous surveys can already be understood as a simplification that helps incorporating the sample design in the process of sample size estimation. However, there are several, more broad-stroked rules of thumb that may help to guide the decision on the appropriate sample size. We discuss the most common and most applicable ones in the following sections. Please keep in mind that these are only rough generalizations. If more/ better information is available, one should use these to get a more appropriate sample size estimation.

Rule of 30 (or 50 or 100)

The so called *rule of 30* states that the sample size for each (sub-) group the researcher wants to examine should equal at least 30 elements. So if, for example, one is interested in the distribution of a variable of interest in dependence of gender, the survey should contain at least 30 male and 30 female participants. If the subgroups are divided by age-class, each of those should contain at least 30 observations. This particular rule is based on the guidelines stipulated by Roscoe (1975). Although that number seems to be arbitrary at first, the reasoning could be determined by the Central Limit Theorem. For most distribu-

tions of sample means it can therefore be assumed that due to the degrees of freedom, a sample size (or subgroup size) of 30 or more is sufficient for the Central Limit Theorem to hold, meaning the distribution approaches normal distribution (Memon et al., 2020).

However, such generalizations should always be taken with care, since there are numerous factors that might also impact the quality of ones (subgroup-) analysis. When, for example, conducting a multivariate regression, another rule of thumb stipulates that the number of observations should be at least 10 times greater than the number of variables in the model (Roscoe, 1975).

Moreover, that generalization does not take variables of interest into account that are highly skewed. In that case, 30 might still not be sufficient. Hence, that number should be taken as a "bare minimum" that may suffice but is not applicable for every scenario. Therefore, one might also encounter the recommendation of 50 or even 100 observations for every subgroup. Even though these numbers seem equally arbitrary, the resulting sample size is estimated more conservatively, which would in term be beneficial for ones analyses.

Sample Size Depending on the Prevalence of "Desired" Observations

Given the previous rule of thumb, it might be rather obvious that the sample size should also consider the prevalence of particular outcomes (or groups) within a particular population. If a subgroup for which analyses are to be made or which should be compared to another subgroup is of a much smaller proportion within the survey's target population, the sample size needs to be larger in order to obtain sufficient observations within the smaller subgroup.

Green's Procedure

The recommendation of Green (1991) is aimed to give a rule of thumb to determine the sample size, when model estimation will be performed using multiple independent variables. His initial suggestion is to aim for a sample size that is

$$n \geq 50 + 8k,$$

where k refers to the number of independent variables used in the model estimation. However, if the researcher is interested in testing individual predictors (rather than the entire model and its coefficient of determination (R^2)), then Green suggests to base the sample size decision on

$$n \geq 104 + k.$$

Moreover, if both is supposed to be tested, the author argues to calculate both sample sizes and decide on the larger one (Green, 1991; Memon et al., 2020).

Please note that when conducting the analyses of subgroups, the sample size refers to the necessary size for each of these groups. Furthermore, it can still be argued that if a larger sample size is achievable, a researcher will detect smaller effect sizes with a greater power (VanVoorhis & Morgan, 2001). Additionally, parameters that are highly skewed may still require to base the model on a larger sample.

Further Rules of Thumb

Other suggestions to determine the sample size are frequently also based on the number of predictors within multivariate models. For instance, the so called *10-times-rule* of Barcley et al. (1995) proposes that within a structural equation model, the sample size should at least be ten times as much as the largest number of structural paths that are directed at a particular path of the models construct (Memon et al., 2020).

Others are based on χ^2-comparison and argue that the minimum number of observations per cell should be at least five, while the minimum size for the entire comparison should be at least 20. A rule of thumb for factor analysis states that at least 300 observations with at least 50 per factor are required (VanVoorhis & Morgan, 2001).

As for multilevel models, a common recommendation is to base the sample size on at least 30 groups containing a minimum number of elements of 30. Alternatively, this rule of thumb has been adjusted to a 50/20 ratio (Memon et al., 2020).

4. Conclusion

Even though sample size calculation is an easy task in theory, there are severe practical challenges that researchers must tackle. Even for a simple random sample, the minimum size heavily relies on population information that are in many cases not available to the researcher. This is even more true for complex samples which show even more unknowns. In this guideline we try to give researchers conducting surveys practical guidance on how to gather information on these parameters and on rules of thumb that can be applied if no other information is available.
Please note that for all the rules of thumb that have been discussed here, one should value them as mere heuristics. The most appropriate of the above mentioned rules of thumb would be the application of prior design effects from earlier studies with the same design. In case of the remaining suggestions, it has to be noticed that they are often based on strong assumption, such as that the variable(s) of interest are normally distributed. A circumstance that is often not met, when surveying a population. Therefore we suggest to apply these rules with caution and use them as a suggested bare minimum sample size. Moreover, we suggest if multiple of these rules apply, to decide on the one that yields the largest sample size.

References

Campbell, M., Grimshaw, J., & Steen, N. (2000). Sample size calculations for cluster randomised trials. *Journal of Health Services Research & Policy*, *5*(1), 12–16. https://doi.org/10.1177/135581960000500105

Champely, S. (2020). *Pwr: Basic functions for power analysis*. Retrieved from https://CRAN.R-project.org/package=pwr

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Gabler, S., & Häder, S. (2015). Stichproben in der Theorie. *Mannheim, GESIS – Leibniz-Institut Für Sozialwissenschaften (GESIS Survey Guidelines)*.

Gabler, S., Häder, S., & Lahiri, P. (1999). A model based justification of kish's formula for design effects for weighting and clustering. *Survey Methodology*, *25*, 105–106.

Gabler, S., Häder, S., & Lynn, P. (2006). Design effects for multiple design samples. *Survey Methodology*, *32*(1), 115–120.

Gelman, A., & Hill, J. (2006). Sample size and power calculations. In *Analytical Methods for Social Research*. *Data analysis using regression and multilevel/hierarchical models* (pp. 437–456). https://doi.org/10.1017/CBO9780511790942.026

Green, S. B. (1991). How many subjects does it take to do a regression analysis. *Multivariate Behavioral Research*, *26*(3), 499–510. https://doi.org/10.1207/s15327906mbr2603_7

Groves, R. M. (2005). *Survey errors and survey costs* (Vol. 581). John Wiley & Sons.

Kalton, G., Brick, J. M., & Lê, T. (2005). Chapter VI estimating components of design effects for use in sample design. In S. F. No. 96. United Nations: Department of Economic & S. A. S. Division (Eds.), *Household sample surveys in developing and transition countries*. Citeseer.

Kish, L. (1987). Weighting in Deft2. *The Survey Statistician*, *17*(1), 26–30.

Koen Beullens, K. D. andCaroline. V., Geert Loosveldt. (2014). *Quality matrix for the european SocialSurvey, round 7*. European Social Survey Round 7.

Lynn, P., Häder, S., & Gabler, S. (2006). Design effects for multiple design samples. *Survey Methodology*, *32*(1), 115–120.

Memon, M. A., Ting, H., Cheah, J.-H., Thurasamy, R., Chuah, F., & Cham, T. H. (2020). Sample size for survey research: Review and recommendations. *Journal of Applied Structural Equation Modeling*, *4*(2), i–xx. https://doi.org/10.47263/jasem.4(2)01

Roscoe, J. T. (1975). *Fundamental research statistics for the behavioral sciences* (2. ed). Retrieved from http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+226871045&sourceid=fbw_bibsonomy

Valliant, R., Dever, J. A., & Kreuter, F. (2013). *Practical tools for designing and weighting survey samples* (Vol. 1). Springer.

Valliant, R., Dever, J. A., & Kreuter, F. (2021). *PracTools: Tools for designing and weighting survey samples*. Retrieved from https://CRAN.R-project.org/package=PracTools

VanVoorhis, C. W., & Morgan, B. L. (2001). Statistical rules of thumb: What we don't want to forget about sample sizes. *Psi Chi Journal of Psychological Research*, 139–141. https://doi.org/10.24839/1089-4136.jn6.4.139